https://doi.org/10.48009/2_iis_2006_218-223

RELIABILITY IN AUTOMATED EVALUATION TOOLS FOR WEB ACCESSIBILITY STANDARDS COMPLIANCE

Ashli M. Molinero, University of Pittsburgh, ashli@pitt.edu
Frederick G. Kohun, Robert Morris University, kohun@rmu.edu

ABSTRACT

This research addressed the problem that neither the design standards nor the automated evaluation tools meet Web designers' requirements for ensuring that there are no accessibility barriers in their Web sites [1, 5]. Different tools arrive at different conclusions when assessing the same Web site for errors. Using Krippendorff's Alpha Reliability Co-efficient (Kr-α) as a measure of inter-reliability, the computer-assisted content analysis tested data from 50 Web sites. These findings support the argument that a human computer interaction approach should be pursued rather than relying on these tools exclusively.

Keywords: Standards Compliance, Web Accessibility, Web Design, Automated Evaluation Tools, Reliability

INTRODUCTION

An accessible Web design is one that allows a person with a disability equal access to the information and content of the site as a person without a disability, regardless of the technology used. The literature regarding Web accessibility explicitly acknowledges that designers should not rely on automated tools alone to validate their sites for compliance. The best way to determine technical accessibility barriers is from a combination of multiple inspection methods including automated tools, usability studies and human review of source code [15, 2, 3, 16, 17, 13, 19]. Considering designers' reliance on the tools in the development of accessible Web sites, the tool's reliability is critical [8].

Achieving Compliance: Automated or Not?

Many automated tools are available to test for compliance to these standards. This is a good indication that the standards were necessary: tools have been developed to support them [12]. When using any automated tool to test for enforceable standards, a user needs to have confidence in the tool's ability to generate accurate, unambiguous test data results [14]. While automated tools have been developed, two or more tools can give different results when assessing the same source code, creating ambiguity for the designer. In addition, the accuracy of the tools is questionable because they have been known to give false validation to sites that still present technical barriers.

In addition, research [5] shows that reliability and reproducibility between the different tools' test data are questionable. Stated otherwise, there is no guarantee two tools will produce the same results.

The Objective

Web designers rely on automated evaluation tools for assessing the accessibility of their sites. Contradicting or discrepant test data from two tools forces a developer to ascertain which tool is generating the accurate report, defeating the purpose of an automated tool. If the automated tools have all been developed to test for the same barriers from objective measurable standards, they should yield consistent results when compared to each other. Therefore, the objective was to determine which United States Section 508 Standards are generating conflicting test data most frequently.

The Purpose and Significance of the Study

The purpose of this project was to establish the inter-reliability of the automated evaluation tools and to subsequently identify those accessibility standards for which automated evaluation is not reaching a sufficient degree of consistency. The implications have significance for multiple reasons. First, designers will benefit from knowing the level of confidence that can be placed on each tool. Second, by identifying those standards that generate the most conflicting test data, designers will know which standards demand a greater depth of understanding. Third, by identifying these standards, we can begin to assess why the process of testing for their compliance has not been successfully automated. Fourth, the United States has set a precedent in adopting the WCAG as a basis for Federal regulations. These standards are being replicated globally; therefore, ensuring they are testable measures is imperative.
Web designers have migrated to automated tools for various reasons, including: difficulty in designing to

the standards due the complexity of the language in which the standards are written [11], unfamiliarity with terms and concepts associated with people with disabilities and assistive technologies [4, 18], and the tools' ability to cost effectively evaluate Web sites with great speed and ease [9]. Research [8] indicates that designers use the tools primarily to ensure accessibility during the design process, but reliability between the tools' test data is questionable [5].

Using multiple evaluation tools is suggested to address vulnerabilities within a single tool [3]. However, discrepancies from different tools, assessing the same data, force a designer to subjectively determine which tool correctly validated the source code. As Brajnik [1] asserts, this by definition is a counter productive task and defeats the purpose of using an "automated" tool.

The United States Section 508 Standards

The U.S. Congress passed an amendment to the Federal Rehabilitation Act of 1973 mandating that any Electronic and Information Technology (EIT) used by the Federal government be accessible to Federal employees or members of the public who may be seeking information from them. In 1998, Section 508 was enacted to include Intra- and Internet technologies [6, 13]. In a letter from Attorney General Janet Reno, all Federal agencies were given two years to establish standards and become compliant with the regulations set forth in Section 508.

METHODOLOGY

The purpose of this research was to investigate the inter-reliability of automated evaluation tools for Web accessibility standards compliance. Test data generated by three automated tools was analyzed for consistency in identifying errors in HTML source code for compliance to U.S. Section 508 Standards. The sample was a purposefully selected group of 50 Web sites. The objective of this study was to identify which standards the automated evaluation tools do not test for compliance reliably or consistently.

Content Analysis Procedures

Though not in the traditional application of a content analysis, the methods used in this research project have attributes consistent with a computer-assisted content analysis. A methodology frequently used in communication research, Krippendorff [10] asserts that "computer assisted content analysis," or "the process of using computers to analyze textual data," has been in practice since the 1950s. A critical component to any content analysis is reliability of the tools employed, and even more so in a computer assisted analysis.

Data Collection Procedures

The first portion of data collection consisted of securing a sample of Web sites to evaluate with the tools for compliance to Section 508. Rather than create a test bed of Web pages that purposely had technical barriers for each standard, the sample in this study was from a variety of "live" Web sites. After the sample was selected, the tests were run to create the reports that were compared for consistency.

Automated Evaluation Tools

The tools selected for this study are automated evaluation tools that have the primary function of assisting Web developers to achieve compliance to standards related to Web accessibility. The HTML source code of a Web site is analyzed for technical barriers that would prevent it from being compliant with Section 508. After analysis, each tool generates reports with the number of errors per Web site. Although the reporting formats differ, each of the automated tools selected generates reports that include information regarding accessibility errors by type (standard), frequency or number of occurrences. Three commercially available tools were used: Watchfire Bobby, LiftNN for Dreamweaver and Ramp. Watchfire Bobby 5.1 was selected because it was the first automated validation tool created. It has, in essence, become the "de-facto standard." LiftNN/g for Macromedia Dreamweaver by UsableNet was selected because it received the highest rating in a report that evaluated six tools for ease of use [7]. Ramp by Deque Systems was selected because it was the newest validation tool at the time and little research was available regarding this product. It should be noted that the default settings for Section 508 compliance tests were selected for each tool.

FINDINGS

Intra-reliability in Automated Evaluation Tools

Intra-reliability is defined as a tool's ability to repeatedly give consistent results for the same test data. To address the issue of intra-reliability, 5 of 50 Web sites, or 10% of the sample, were analyzed twice to establish that each tool could reproduce the same results when assessing the same HTML source code more than once. This analysis was conducted first to ensure the tools were stable and could be used in the study.

The data in Table 1 demonstrates just how much variation comes into play for Web designers when using multiple tools to evaluate their designs for compliance to Section 508 Standards. If the Web sites were being tested for compliance to objective standards, then the results for an individual Web site would be consistent across each row, with minimal discrepancies. This is not the case, however. The nearest the tools came to agreement on a Web site was site 48, where Lift and Ramp both identified a total of 8 errors, and Bobby found a total of 6 errors. This case demonstrates that the tools are capable of achieving a level of agreement with relatively high inter-reliability. However, Web site 10 is an example of the lowest level of agreement between tools in these results. In the source code from Web site 10, Lift found 13 errors, Bobby found 631 errors, and Ramp found 49 errors.

Table 1. Total Number of Errors that were Identified in each Web site by each Tool

Standard	Lift	Bobby	Ramp	Total
508 (a)	2628	1741	757	5126
508 (b)	26	0	0	26
508 (c)	47	3753	0	3800
508 (d)	93	37	4	134
508 (e)	1	2	0	3
508 (f)	1	1	1	3
508 (g)	29	505	0	534
508 (h)	21	1195	0	1216
508 (i)	12	23	24	59
508 (j)	101	46	4	151
508 (k)	41	3	0	44
508 (l)	815	40	152	1007
508 (m)	8	10	5	23
508 (n)	228	191	205	624
508 (o)	206	36	321	563
508 (p)	2	48	3	53
Total	4259	7631	1476	13366

The discrepancies in these results represent the amount of subjectivity a Web designer would have to use in determining which tool is correct, with each discrepancy requiring a human judgment call. Table 2 demonstrates that each standard was tested at least once in the study. The table shows the cumulative number of errors each tool found per standard out of the entire sample.

Table 2. Number of Errors Reported by Automated Evaluation Tools, for each Standard

Web site	Lift	Bobby	Ramp	Web site	Lift	Bobby	Ramp
1	44	143	55	26	39	58	22
2	32	51	--	27	240	375	82
3	22	12	3	28	197	148	92
4	228	185	35	29	--	162	65
5	40	29	22	30	127	263	70
6	155	622	16	31	69	92	26
7	--	--	--	32	46	235	37
8	62	45	19	33	39	75	11
9	44	47	22	34	137	63	14
10	130	631	49	35	64	63	12
11	9	5	0	36	46	186	23
12	91	103	28	37	84	128	7
13	29	66	1	38	72	77	8
14	--	29	12	39	59	167	25
15	54	73	12	40	93	149	33
16	54	136	25	41	223	337	47
17	25	32	5	42	128	233	79
18	35	7	6	43	4	5	1
19	28	53	10	44	227	349	55
20	65	81	34	45	27	17	9
21	71	228	32	46	73	118	9
22	95	78	46	47	225	502	44
23	115	101	18	48	8	6	8
24	51	107	22	49	246	421	49
25	56	21	10	50	251	517	166

This table also demonstrates the differences in the raw data for the total number of errors found by the tools. From the table, Bobby reported a substantially higher number of errors than either Lift or Ramp

Nominal Level Analysis

The results yielded reliability alphas (Kr-α) ranging from -.3173 for Standard (k) which addresses the use of a text-only alternative Web page, to 1.0 for Standard (f) which addresses the use of frames. Table 3 lists the results from these computations.

The findings demonstrate that there are substantial discrepancies in the results between the automated evaluation tools Bobby, Lift and Ramp when testing Web sites for compliance to Section 508 Standards.

Table 3. Results from Computing the Kr-α Reliability in Microsoft Excel.

Standard	Kr-α	Decision
508 (a)	0.1319	not reliable
508 (b)	-0.0301	not reliable
508 (c)	-0.4588	possible systematic disagreement
508 (d)	-0.1211	not reliable
508 (e)	0.3235	not reliable
508 (f)	1	reliable
508 (g)	-0.2560	possible systematic disagreement
508 (h)	-0.3173	possible systematic disagreement
508 (i)	0.7833	reliable
508 (j)	-0.2106	possible systematic disagreement
508 (k)	-0.3912	possible systematic disagreement
508 (l)	0.0734	not reliable
508 (m)	0.5338	possibly reliable
508 (n)	0.8301	reliable
508 (o)	0.3538	not reliable
508 (p)	-0.3508	possible systematic disagreement

To resolve these discrepancies, Web developers must determine by making subjective judgments as to which tools' results are most accurate. Although individually the tools are stable, when the test data reports are compared to each other, they disagree more often than not regarding what constitutes an error in the source code of a Web site.

MAPPING BACK TO THE STANDARDS

Table 4 summarizes the decisions that were made in the previous section regarding objective and subjective components in written requirements of the Section 508 Standards. Of the 16 Section 508 Standards, only four were identified as purely objective standards: (f), (i), (m) and (n). These same four standards had the highest Kr-α levels in the analysis, 3 of which were at .783 or above: (f), (i), and (n). The fourth (m) was at .533 at the ratio level and .500 at the nominal level. The results from these three tools indicate that automated testing is not reliable for compliance to requirements set forth in Section 508 of the United States Rehabilitation Act.

Table 4. Objectivity, Subjectivity and Reliability in Section 508 Standards

508 Standard	Objective	Subjective	Reliable Kr-α
(a)	●	●	
(b)	●	●	
(c)		●	
(d)		●	
(e)	●	●	
(f)	●		●
(g)	●	●	
(h)	●	●	
(i)	●		●
(j)	●	○	
(k)	●	●	
(l)	●	●	
(m)	●		○
(n)	●		●
(o)	●	●	
(p)	●	●	

Accessible Web Design

The most reliably detected errors were those related to standards related to the technical function or architectural structure of a Web page, including the use of frames, form elements and image maps. On a positive note, the tools are able to reliably detect errors related to form elements.

Automated Tools Supporting Standards Compliance

In relation to the literature regarding the automated tools intended to support the standards, several assertions can be made. The problem in reliably detecting errors for the compliance lies within the subjective components of the standards. The

subjectivity in these standards stems from issues pertaining to design elements and human perception or processing of information on a Web page.

CONCLUSION

The subjectivity in the standards is problematic for creating reliable automated validation tools. Enforcing a plan that suggests using tools that are known to be flawed merely perpetuates the problems for both designers and people with disabilities. At this point in time, eliminating the technical barriers on the Internet requires a much more pragmatic approach to Web accessibility. Until the technical standards for Web accessibility can be readdressed, an approach based on skill building for developers rather than automated tools may be more advantageous for both Web developers and people with disabilities.

REFERENCES

1. Brajnik, G. (2001). Towards valid quality models for Web sites. Paper presented at the 7th Conference on Human Factors and the Web, Madison, WI. Retrieved December 15, 2003 from http://www.dmi.uniud.it/~giorgio/papers/hfWeb01.html.
2. Brewer, J. (2001). Access to the World Wide Web: technical and policy perspectives. In Wolfgang F.E. Preiser & Elaine Ostroff (Eds.). *Universal design handbook* (pp.66.1-66.13) New York: McGraw Hill.
3. Brewer, J. & Letourneau, C. (Eds.) (2002). Evaluating Web sites for accessibility. World Wide Web Consortium Web Accessibility Initiative. Retrieved June 1, 2003 from http://www.w3.org/WAI/eval/
4. Colwell, C., & Petrie, H.(1999). Evaluation of guidelines for designing accessible Web content. In C. Buhler & H. Knops (Eds) *Assistive technology on the threshold of the new millennium* (IOS press).
5. Diaper, D. & Worman, L. (2003). Two falls out of three in the automated accessibility assessment of world wide Web sites: a-prompt v. bobby. [Electronic Version]
6. Federal Register (2000, December 21). Electronic and information technology accessibility standards. Retrieved, January 2, 2002 from http://www.accessboard.gov/sec508/508standards.htm
7. Graves, S. (2001). Check sites for 508 with audit-edit tools. Government Computer News, Retrieved April 14, 2003 from http://www.gcn.com/cgi-bin/udt/im.display.printable?client.id=gcn2&story.id=16783.
8. Ivory, M. Y. & Chevalier, A. (2002). A study of automated Web site evaluation tools. Technical Report UW-CSE-02-10-01. [Electronic Version].
9. Killam, B., & Holland, B. (2001). Position paper on the suitability of task automated utilities for testing Web accessibility compliance. Usability Professionals' Association Conference, 2001 Retrieved April 17, 2003 from http://www.upassoc.org/conf2001/reg/program/workshops/w6.html.
10. Krippendorf, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills, CA: Sage Publications, Inc.
11. Lindenberg, J., & Neerinex, M.A. (1999, August). The need for a "universal accessibility" engineering tool. *Proceedings, Interact '99 workshop: Making designers aware of existing guidelines for accessibility*.
12. Morell, J.A., & Stewart, S. (1996, March). Standards development for information technology: best practices for the United States [Electronic Version]. StandardView, *4*(1), 42-51 ACM.
13. Mueller, J.P. (2003). *Accessibility for everybody: understanding the section 508 accessibility requirements*. New York: Apress.
14. Ramamoorthy, C.V. & Ho, S.F. (1975). Testing large software with automated software evaluation systems.
15. Rowan, M., Gregor, P. Sloan, D., & Booth, P. (2000, November). Evaluating Web resoucres for disability access [Electronic Version]. *Assests '00*. 13-15, ACM.
16. Slatin, J.M. & Rush, S. (2003). *Maximum accessibility: making your Web site more usable for everyone*. Pearson Education. New York: Addison-Wesley.
17. Thatcher, J., Bohman, P., Burks, M., Henry, S.L., Regan, B., Swierenga, S. Urban, M.D., & Wadell, C.D. (2002). *Constructing accessible Web sites*. Birmingham, U.K.: Glaushauss Ltd.
18. Velasco, C., & Verelest, T. (1999, August). Raising awareness among designers of accessibility issues. *Proceedings, Interact '99 workshop: Making designers aware of existing guidelines for accessibility.*
19. Zeldman, J. (2003). *Designing with Web Standards*. Boston: New Riders.